Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **85**, 1659–1679
https://doi.org/10.1093/jrsssb/qkad070
Advance access publication 20 July 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology

Core shrinkage covariance estimation for matrix-variate data

Peter Hoff[1], Andrew McCormack[1] and Anru R. Zhang[2]

[1]Department of Statistical Science, Duke University, Durham, USA
[2]Department of Biostatistics and Bioinformatics, Duke University, Durham, USA

Address for correspondence: Peter Hoff, Department of Statistical Science, Box 90251, Durham, NC 27708-0251, USA.
Email: peter.hoff@duke.edu

Abstract

A separable covariance model can describe the among-row and among-column correlations of a random matrix and permits likelihood-based inference with a very small sample size. However, if the assumption of separability is not met, data analysis with a separable model may misrepresent important dependence patterns in the data. As a compromise between separable and unstructured covariance estimation, we decompose a covariance matrix into a separable component and a complementary 'core' covariance matrix. This decomposition defines a new covariance matrix decomposition that makes use of the parsimony and interpretability of a separable covariance model, yet fully describes covariance matrices that are non-separable. This decomposition motivates a new type of shrinkage estimator, obtained by appropriately shrinking the core of the sample covariance matrix, that adapts to the degree of separability of the population covariance matrix.

Keywords: decorrelation, equivariance, Kronecker product, matrix decomposition, matrix square root, whitening

1 Introduction

Many modern data sets include matrix-variate data, that is, a sample of n matrices $Y_1, \ldots, Y_n$ having a common dimension $p_1 \times p_2$. Examples of such data sets include collections of images, networks, gene by tissue expression arrays, and multivariate time series, among others. One approach to the analysis of such data is to first vectorise each data matrix by stacking columns, and then proceed with a method that is appropriate for generic multivariate data. For example, if $Y_1, \ldots, Y_n$ is a random sample from a population of mean-zero matrices, the population covariance could be estimated by the sample covariance $S = \sum_{i=1}^{n} y_i y_i^\top / n$, where for $i = 1, \ldots, n$, y_i is the vector of length $p = p_1 \times p_2$ obtained by vectorising Y_i.

However, in many applications the sample size n is insufficient for such unstructured estimates to be statistically stable. For example, even though p_1 and p_2 might be of moderate magnitude individually, a sample size of $n \geq p_1 p_2$ is necessary for S to be non-singular, and for the likelihood corresponding to a normal model to be bounded. Furthermore, even if the sample size is sufficient for estimation, an unstructured estimate such as S may be difficult to interpret, as it is not expressed in terms of conceptually simple row factors or column factors.

For these reasons, covariance models that are based on the matrix structure of the data have been developed. Most popular are the separable or Kronecker-structured covariance models that assume the $p \times p$ population covariance matrix is the Kronecker product of two smaller covariance matrices of dimension $p_1 \times p_1$ and $p_2 \times p_2$, representing across-row and across-column covariance, respectively. In particular, the separable covariance model for normally distributed data (Dawid 1981) has been used for a wide variety of applications including environmental monitoring (Mardia & Goodall 1993), signal processing (Werner et al. 2008), image analysis (Zhang & Schneider 2010), gene expression data (Yin & Li 2012), radar detection (Greenewald et al. 2016), and many others.

Received: December 10, 2022. **Revised:** May 22, 2023. **Accepted:** June 19, 2023

In addition to its interpretability, a separable covariance model is appealing because of its statistical stability, which is a result of its parsimony as compared to an unstructured covariance model. Remarkably, the maximum likelihood estimator (MLE) in the separable normal model exists uniquely for any sample size n larger than $p_1/p_2 + p_2/p_1$ (Derksen & Makam 2021; Drton et al. 2021; Roś et al. 2016; Soloveychik & Trushin 2016). This is in contrast to a sample size requirement of $n \geq p_1 p_2$ in a normal model with an unstructured covariance. However, the appropriateness of a separable covariance estimator depends on the extent to which the population covariance is truly separable. If the population covariance is not separable, a separable estimate would give an incomplete summary of the statistical dependencies in the data, and could lead to poor performance of statistical procedures, such as generalised least-squares or quadratic discriminant analysis (QDA), that rely on an accurate estimate of the population covariance. These and other concerns about the appropriateness of the separability assumption have been raised by M. L. Stein (2005) and Rougier (2017), specifically in the context of random spatio-temporal processes. To address these concerns, Masak and Panaretos (2022) and Masak et al. (2023) have proposed generalisations of the class of separable covariance operators for functional data analysis with two-dimensional domains (e.g. space and time). The first of these is based on an approximation of an arbitrary positive-definite covariance operator by a sum of separable matrices. The second of these assumes the covariance operator is the sum of two positive-definite operators, one of which is separable and the other is banded, where the banding is determined by the metrics of each of the two domains.

Instead of approximating a covariance matrix, the approach we take in this article is to represent a covariance matrix in terms of a single separable component and a complementary non-separable component. This representation provides a new parametrisation of the space of covariance matrices that makes use of the parsimony and interpretability of a separable covariance model, yet fully describes covariance matrices that are non-separable. This parametrisation motivates a type of covariance shrinkage estimator that can have a risk that is comparable to that of a separable estimator when the population covariance is truly separable, and otherwise has a lower risk than both the separable and unstructured estimators.

In the next section, we define the Kronecker covariance and core covariance of an arbitrary $p_1 p_2 \times p_1 p_2$ covariance matrix. We show that the space of all $p_1 p_2 \times p_1 p_2$ covariance matrices can be identifiably parametrised by the product space of Kronecker and core covariance matrices using this Kronecker–core decomposition (KCD). The Kronecker covariance provides a precise and interpretable definition of the 'separable part' of a non-separable covariance matrix, which can facilitate interpretation of covariance estimates in data analysis situations where the population covariance matrix is likely not separable. From a geometric perspective, the core covariance indicates where a given covariance matrix is along a sub-space that is orthogonal to a tangent space to the set of separable covariance matrices. This motivates a class of core shrinkage estimators, developed in Section 3, that are obtained by linearly shrinking the core of the sample covariance matrix towards the identity matrix. This is equivalent to shrinking the sample covariance matrix S towards its separable part $k(S)$, resulting in a simple estimator $\hat{\Sigma}$ of the form $\hat{\Sigma} = (1-w)S + wk(S)$ for some $w \in [0, 1]$. This shrinkage estimator can be positive-definite even when the sample size is much smaller than the dimension of the covariance matrix. We use an empirical Bayes approach to estimate an appropriate amount of shrinkage w from the data, and show that the resulting core shrinkage estimator (CSE) is consistent. In a simulation study in Section 4.1, we show that the loss of the proposed core shrinkage estimator can be very close to that of an oracle Bayes estimator, and lower than that of both the separable and unstructured MLEs across a variety of conditions. In Section 4.2, we use CSEs as inputs into a QDA for a speech recognition task. We observe that classifications using CSEs have lower out-of-sample misclassification rates than those using separable or unstructured MLEs. A discussion of directions for further research follows in Section 5. Proofs of mathematical results are provided in Appendix A.

2 Kronecker and core covariances

2.1 The Kronecker covariance of a random matrix

Let Y be a mean-zero random matrix taking values in $\mathbb{R}^{p_1 \times p_2}$, and let $y \in \mathbb{R}^p$ be its vectorisation, so that $p = p_1 p_2$. We define the covariance matrix $\mathrm{Var}[Y]$ of Y to be the $p \times p$ matrix

$\Sigma = \mathrm{Var}[y] = \mathrm{E}[yy^\top]$, which we assume to be non-singular and therefore a member of the set $\mathcal{S}_p^+$ of positive-definite $p \times p$ matrices. Recall that Σ is *Kronecker separable*, or simply *separable*, if it can be expressed as $\Sigma = \Sigma_2 \otimes \Sigma_1$ for some matrices $\Sigma_1 \in \mathcal{S}_{p_1}^+$, $\Sigma_2 \in \mathcal{S}_{p_2}^+$, where '$\otimes$' is the Kronecker product (Dutilleul 1999; Srivastava et al. 2008) In this case, the matrices Σ_1, Σ_2 (or matrices $c\Sigma_1, \Sigma_2/c$ for any $c > 0$) are often referred to as the row covariance and column covariance of Y, respectively. For example, the covariance of the p_2 random variables in a common row of Y is proportional to Σ_2, and so Σ_2 represents the covariances of the elements of Y across its columns.

Let $\mathcal{S}_{p_1,p_2}^+ = \{\Sigma_2 \otimes \Sigma_1 : \Sigma_1 \in \mathcal{S}_{p_1}^+, \Sigma_2 \in \mathcal{S}_{p_2}^+\} \subset \mathcal{S}_p^+$ be the set of separable covariance matrices for given values of p_1 and p_2. A separable covariance model is a collection of probability distributions for Y for which it is assumed that $\mathrm{Var}[Y] \in \mathcal{S}_{p_1,p_2}^+$. The most widely used separable model is the separable normal model, or 'matrix normal' model (Dawid 1981), which specifies that $Y \sim N_{p_1 \times p_2}(0, \Sigma_2 \otimes \Sigma_1)$ for unknown $\Sigma_2 \otimes \Sigma_1 \in \mathcal{S}_{p_1,p_2}^+$. A separable covariance model can be thought of as a bilinear transformation model: Let Z be a $p_1 \times p_2$ mean-zero random matrix with $\mathrm{Var}[Z] = I_p$, and let $Y = A_1 Z A_2^\top$ for non-singular matrices $A_1 \in \mathbb{R}^{p_1 \times p_1}$, $A_2 \in \mathbb{R}^{p_2 \times p_2}$. Then $\mathrm{Var}[Y] = A_2 A_2^\top \otimes A_1 A_1^\top$, and the range of $\mathrm{Var}[Y]$ over all such matrices A_1, A_2 is exactly equal to $\mathcal{S}_{p_1,p_2}^+$. More generally, separability is preserved under row and column transformations of Y: If $\mathrm{Var}[Y] = \Sigma_2 \otimes \Sigma_1$, then

$$\begin{aligned} \mathrm{Var}[A_1 Y A_2^\top] \equiv \mathrm{Var}[(A_2 \otimes A_1)y] &= (A_2 \otimes A_1)\mathrm{Var}[y](A_2 \otimes A_1)^\top \\ &= (A_2 \otimes A_1)(\Sigma_2 \otimes \Sigma_1)(A_2 \otimes A_1)^\top \\ &= (A_2 \Sigma_2 A_2^\top) \otimes (A_1 \Sigma_1 A_1^\top). \end{aligned} \tag{1}$$

In the language of group theory, let $GL_{p_1,p_2} = \{A_2 \otimes A_1 : A_1 \in GL_{p_1}, A_2 \in GL_{p_2}\}$ be the separable sub-group of the general linear group GL_p of non-singular $p \times p$ matrices. The transformation in equation (1) from $\mathrm{Var}[Y]$ to $\mathrm{Var}[A_1 Y A_2^\top]$ defines a transitive group action of GL_{p_1,p_2} on $\mathcal{S}_{p_1,p_2}^+$. The group structure of the separable normal model and related tensor normal models has been exploited to develop methods for statistical estimation (Gerard & Hoff 2015) and testing (Gerard & Hoff 2016; Hoff 2016).

Even if $\mathrm{Var}[Y]$ is not separable, it still may be of interest to define some notion of row covariance and column covariance for Y. To this end, we identify a separable covariance matrix $K \in \mathcal{S}_{p_1,p_2}^+$ that summarises the row and column covariance of Y when $\mathrm{Var}[Y]$ is an arbitrary covariance matrix $\Sigma \in \mathcal{S}_p^+$:

Definition 1 Let $\mathrm{E}[Y] = 0$ and $\mathrm{Var}[Y] = \Sigma \in \mathcal{S}_p^+$. The Kronecker covariance of Σ is $k(\Sigma) = \Sigma_2 \otimes \Sigma_1$, where (Σ_1, Σ_2) are any matrices in $\mathcal{S}_{p_1}^+ \times \mathcal{S}_{p_2}^+$ that satisfy

$$\begin{aligned} \Sigma_1 &= \mathrm{E}[Y \Sigma_2^{-1} Y^\top]/p_2 \\ \Sigma_2 &= \mathrm{E}[Y^\top \Sigma_1^{-1} Y]/p_1. \end{aligned} \tag{2}$$

Matrices Σ_1 and Σ_2 that solve equation (2) are weighted averages of across-row and across-column covariance matrices of whitened versions of Y. For example, Σ_1 is obtained from Y by first whitening across its columns by Σ_2.

Solutions to equation (2) exist for all $\Sigma \in \mathcal{S}_p^+$, and all solutions have the same Kronecker product, so the Kronecker covariance function $k : \mathcal{S}_p^+ \to \mathcal{S}_{p_1,p_2}^+$ is well defined. The existence and uniqueness of a $\Sigma_2 \otimes \Sigma_1$ that satisfies equation (2) follow from existing results for the separable normal model, and the following alternative definition of $k(\Sigma)$ as the element of $\mathcal{S}_{p_1,p_2}^+$ that is closest to Σ in terms of a standard divergence function:

Proposition 1 (Σ_1, Σ_2) is a solution to equation (2) if and only if $\Sigma_2 \otimes \Sigma_1$ minimises $d(K : \Sigma) = \ln |K| + \mathrm{trace}(K^{-1}\Sigma)$ over $K \in \mathcal{S}_{p_1,p_2}^+$.

The divergence function $d(K : \Sigma)$ is related to Stein's loss for covariance estimation and to the Kullback-Leibler divergence between two normal distributions. Specifically, $k(\Sigma)$ is the covariance

matrix of the separable normal distribution that minimises the Kullback-Leibler divergence to the $N_{p_1\times p_2}(0,\Sigma)$ distribution. This means that if $Y_1,\ \ldots,\ Y_n \sim$ i.i.d. $N_{p_1\times p_2}(0,\Sigma)$ then the MLE of $\Sigma_2\otimes\Sigma_1$ under the potentially mis-specified model $Y_1,\ \ldots,\ Y_n \sim$ i.i.d. $N_{p_1\times p_2}(0,\Sigma_2\otimes\Sigma_1)$ converges in probability to $k(\Sigma)$ as $n\to\infty$ (Huber 1967). In the language of mis-specified models, $k(\Sigma)$ is the 'pseudo-true' parameter under the separable normal model in the case that Σ is not necessarily separable.

That the minimiser of the divergence function is unique follows from uniqueness results for the MLE in the separable normal model. The MLE for this model is obtained by minimising over $\Sigma_2\otimes\Sigma_1\in\mathcal{S}^+_{p_1,p_2}$ the scaled log-likelihood

$$(-2/n)\times\ln p(Y_1,\ \ldots,\ Y_n|\Sigma_2\otimes\Sigma_1)=\ln|\Sigma_2\otimes\Sigma_1|+\mathrm{trace}((\Sigma_2\otimes\Sigma_1)^{-1}S)+p\ln 2\pi,$$

where $S=\sum_{i=1}^n y_i y_i^\top/n$ is the sample covariance matrix. Clearly, the conditions on S for there to exist a unique MLE of $\Sigma_2\otimes\Sigma_1$ are the same as those on Σ for there to exist a unique minimiser of $d(K:\Sigma)$ over $K\in\mathcal{S}^+_{p_1,p_2}$. In particular, $k(S)$ is the MLE of $\Sigma_2\otimes\Sigma_1$ under the separable normal model when S is the sample covariance matrix. Srivastava et al. (2008) show that this MLE exists uniquely if S is strictly positive-definite, which implies that $k(\Sigma)$ exists uniquely for any $\Sigma\in\mathcal{S}^+_p$. We note that solutions may also exist uniquely when S, or analogously Σ, is singular (Derksen & Makam 2021; Drton et al. 2021; Soloveychik & Trushin 2016).

Numerical methods for finding the separable normal MLE may be used to compute the Kronecker covariance function. As shown in Dutilleul (1999), $\hat\Sigma_2\otimes\hat\Sigma_1$ is an MLE of $\Sigma_2\otimes\Sigma_1$ if $(\hat\Sigma_1,\hat\Sigma_2)$ satisfy

$$\begin{aligned}\left(\sum_{i=1}^n Y_i\hat\Sigma_2^{-1}Y_i^\top/n\right)\Big/p_2&=\hat\Sigma_1\\ \left(\sum_{i=1}^n Y_i^\top\hat\Sigma_1^{-1}Y_i/n\right)\Big/p_1&=\hat\Sigma_2.\end{aligned}\tag{3}$$

Dutilleul also provided a block co-ordinate descent algorithm that converges to the MLE when it exists uniquely. Because this system of equations is analogous to the system (2) that define $k(\Sigma)$, Dutilleul's algorithm may be implemented to numerically compute the Kronecker covariance $k(\Sigma)$ of any $\Sigma\in\mathcal{S}^+_p$. In this context, given a starting value $\Sigma_2\in\mathcal{S}^+_{p_2}$, the algorithm is to iterate the following steps until a convergence criteria is met:

1. Set $\Sigma_1=\mathrm{E}[Y\Sigma_2^{-1}Y^\top]/p_2$;
2. Set $\Sigma_2=\mathrm{E}[Y^\top\Sigma_1^{-1}Y]/p_1$.

An algorithm to compute $k(\Sigma)$ is provided in the replication material for this article.

An important property of the Kronecker covariance function is how it is affected by transformations of Σ, or equivalently, of Y. Recall that if Y has a separable covariance $\Sigma_2\otimes\Sigma_1$, then $A_1YA_2^\top$ has separable covariance $(A_2\Sigma_2A_2^\top)\otimes(A_1\Sigma_1A_1^\top)$, and so in this sense a linear transformation across the rows of Y changes the row covariance and not the column covariance, and analogously for a column transformation. The following result shows that the Kronecker covariance function transforms in the same way, even if the covariance matrix of Y is not separable:

Proposition 2 For $A_2\otimes A_1\in GL_{p_1,p_2}$ and $\Sigma\in\mathcal{S}^+_p$ with $k(\Sigma)=\Sigma_2\otimes\Sigma_1$,

$$\begin{aligned}k((A_2\otimes A_1)\Sigma(A_2\otimes A_1)^\top)&=(A_2\otimes A_1)k(\Sigma)(A_2\otimes A_1)^\top.\\&=(A_2\Sigma_2A_2^\top)\otimes(A_1\Sigma_1A_1^\top).\end{aligned}$$

From the perspective of group theory, the group action of GL_{p_1,p_2} on $\mathbb{R}^{p_1\times p_2}$ defined by $Y\mapsto A_1YA_2^\top$ induces a group action of GL_{p_1,p_2} on $\mathcal{S}^+_p$ given by $\Sigma\mapsto(A_2\otimes A_1)\Sigma(A_2\otimes A_1)^\top$. The result

is that the Kronecker covariance function k is equivariant with respect to this group action—the Kronecker covariance of the separably transformed Σ is the separably transformed Kronecker covariance of Σ. This property will be used throughout the remainder of this article. Additional properties of the Kronecker covariance function include the following:

Corollary 1
1. $k(I_p) = I_p$.
2. If $\Sigma \in \mathcal{S}^+_{p_1,p_2}$ then $k(\Sigma) = \Sigma$.
3. For $a > 0$, $k(a\Sigma) = ak(\Sigma)$.
4. If Σ is diagonal then $k(\Sigma)$ is diagonal.

The third item indicates that k is a scale-equivariant function. As a result, the shrinkage estimator we propose in Section 3 will be scale-equivariant.

2.2 The Kronecker–core parametrisation and decomposition

The Kronecker covariance function k defined above is a surjection from $\mathcal{S}^+_p$ to $\mathcal{S}^+_{p_1\times p_2}$ that describes the row covariance and column covariance of an arbitrary element of $\mathcal{S}^+_p$. We now use this function to define, for each $\Sigma \in \mathcal{S}^+_p$, a 'core' covariance matrix $c(\Sigma)$ that is complementary to $k(\Sigma)$ in that the core lacks across-row and across-column covariance in some sense. We then show that, taken together, the product space of separable and core covariance matrices identifiably parametrises $\mathcal{S}^+_p$.

A core covariance of $\Sigma \in \mathcal{S}^+_p$ is obtained by applying a transformation to Σ that whitens its Kronecker covariance. Specifically, let $H = H_2 \otimes H_1$ be any matrix in GL_{p_1,p_2} such that $HH^\top = k(\Sigma)$. By the equivariance of k, we have

$$\begin{aligned} k(H^{-1}\Sigma H^{-\top}) &= H^{-1}k(\Sigma)H^{-\top} \\ &= H^{-1}HH^\top H^{-\top} = I_p. \end{aligned}$$

We define Kronecker-whitened versions of Σ as follows:

Definition 2 Let $H = H_2 \otimes H_1 \in GL_{p_1,p_2}$ satisfy $HH^\top = k(\Sigma)$. Then the matrix C given by $C = H^{-1}\Sigma H^{-\top}$ is a core of Σ.

We call the matrix $C = H^{-1}\Sigma H^{-\top}$ a core of Σ because the four-way tensor $\tilde{\Sigma} \in \mathbb{R}^{p_1\times p_2\times p_1\times p_2}$ with entries corresponding to Σ may be expressed in terms of C, H_1 and H_2 via the multilinear operation

$$\tilde{\Sigma} = \tilde{C} \times \{H_1, H_2, H_1, H_2\},$$

where '$\times$' is the multilinear product and $\tilde{C}$ is the four-way tensor corresponding to C. Equivalently, we have $\mathrm{vec}(\Sigma) = (H_2 \otimes H_1 \otimes H_2 \otimes H_1)\mathrm{vec}(C)$. In the context of Tucker products, the tensor that gets multiplied along each mode by a matrix is called the 'core'.

There are multiple cores for a given Σ, as there are multiple separable matrices H for which $HH^\top = k(\Sigma)$. Conversely, a core of Σ is also a core of $H\Sigma H^\top$ for any $H \in GL_{p_1,p_2}$. More generally, we say that a covariance matrix $C \in \mathcal{S}^+_p$ is a core covariance matrix if it is the core of some $\Sigma \in \mathcal{S}^+_p$, and so any core covariance matrix satisfies $k(C) = I_p$. Furthermore, suppose $C \in \mathcal{S}^+_p$ satisfies $k(C) = I_p$. Then C is a core of any covariance matrix $HCH^\top$ for $H \in \mathcal{S}^+_{p_1,p_2}$. As such, for a given p_1 and p_2, we define the set of core matrices as follows:

Definition 3 For a given p_1 and p_2 with $p_1 \times p_2 = p$, the set of core covariance matrices is $\mathcal{C}^+_{p_1,p_2} = \{C \in \mathcal{S}^+_p : k(C) = I_p\}$.

The condition $k(C) = I_p$ defining $\mathcal{C}^+_{p_1,p_2}$ can alternatively be expressed as follows:

Proposition 3 Let Y have covariance matrix $C \in \mathcal{S}^+_p$, and let $\tilde{C}$ be the $p_1 \times p_2 \times p_1 \times p_2$ tensor where $\tilde{C}_{i,j,i',j'} = \mathrm{Cov}[Y_{i,j}, Y_{i',j'}]$. Then $k(C) = I_p$ if and only if

$$\mathrm{E}[YY^{\top}]/p_2 \equiv \sum_{j=1}^{p_2} \tilde{C}_{\cdot,j,\cdot,j}/p_2 = I_{p_1}$$

$$\mathrm{E}[Y^{\top}Y]/p_1 \equiv \sum_{i=1}^{p_1} \tilde{C}_{i,\cdot,i,\cdot}/p_1 = I_{p_2}.$$

So for a core covariance matrix, the across-column average of the across-row covariance is the identity matrix, and analogously for the across-column covariance. Intuitively, a core covariance has no across-row or across-column correlation or heteroscedasticity, on average.

From the proposition we see that $\mathcal{C}^{+}_{p_1,p_2}$ is defined by a system of linear constraints, and that $\mathrm{trace}(C) = p$ for any core covariance matrix C, so $\mathcal{C}^{+}_{p_1,p_2}$ is a compact convex subset of $\mathcal{S}^{+}_{p}$. Additionally, the core covariances $\mathcal{C}^{+}_{p_1,p_2}$ and the separable covariances $\mathcal{S}^{+}_{p_1,p_2}$ are nearly non-overlapping: If C is a core matrix then $k(C) = I_p$, and if C is separable, then $k(C) = C$ by Corollary 1. Therefore, if C is core and separable, then $C = I_p$. Thus, $\mathcal{S}^{+}_{p_1,p_2} \cap \mathcal{C}^{+}_{p_1,p_2} = I_p$.

For every $\Sigma \in \mathcal{S}^{+}_{p}$ there is a core matrix $C \in \mathcal{C}^{+}_{p_1,p_2}$ and separable matrix $K \in \mathcal{S}^{+}_{p_1,p_2}$ such that $\Sigma = HCH^{\top}$ for some separable $H \in GL_{p_1,p_2}$ such that $K = HH^{\top}$. Conversely, to every $C \in \mathcal{C}^{+}_{p_1,p_2}$ and $K \in \mathcal{S}^{+}_{p_1,p_2}$ we can define an element of $\mathcal{S}^{+}_{p}$ as $HCH^{\top}$, where $H \in GL_{p_1,p_2}$ and $K = HH^{\top}$. This suggests that there is a bijection between $\mathcal{S}^{+}_{p}$ and $\mathcal{S}^{+}_{p_1,p_2} \times \mathcal{C}^{+}_{p_1,p_2}$. In fact, there are many such bijections, including one for each way to define a separable matrix square root H of K, or equivalently, one for each way to define a row and column whitening matrix from K. To specify a particular bijection, we need to specify a separable square root function.

Definition 4 Let $\mathcal{H}$ be a subset of GL_{p_1,p_2} such that the function $s : \mathcal{H} \to \mathcal{S}^{+}_{p_1,p_2}$ defined by $s(H) = HH^{\top}$ is a bijection. Then $h = s^{-1}$ is a separable matrix square root function.

Essentially, a separable square root function is defined by a set of separable matrices $\mathcal{H}$ with unique cross-products, the set of which equals the set of separable covariance matrices. The defining feature of such a function is that $h(HH^{\top}) = H$ for $H \in \mathcal{H}$. Two examples include the following:

- Symmetric square root: $h(\Sigma_2 \otimes \Sigma_1) = \Sigma_2^{1/2} \otimes \Sigma_1^{1/2}$, where $\Sigma_j^{1/2}$ is the symmetric square root of Σ_j, $j \in \{1, 2\}$.
- Cholesky square root: $h(\Sigma_2 \otimes \Sigma_1) = L_2 \otimes L_1$, where $L_j L_j^{\top}$ is the lower triangular Cholesky factorisation of Σ_j, $j \in \{1, 2\}$.

A non-example would be GL_{p_1,p_2}: Whilst the set of cross-products of this set is equal to $\mathcal{S}^{+}_{p_1,p_2}$, elements of the set do not have unique cross-products.

For a given separable square root function h we define the core covariance function $c : \mathcal{S}^{+}_{p} \to \mathcal{C}^{+}_{p_1,p_2}$ as $c(\Sigma) = H^{-1}\Sigma H^{-\top}$, where $H = h(k(\Sigma))$. Since the core represents 'non-separable' covariance, we would hope the core function to be invariant to bilinear transformations of the form $\Sigma \mapsto (A_2 \otimes A_1)\Sigma(A_2 \otimes A_1)^{\top}$ that induce separable covariance. This property partly holds:

Proposition 4 Let $\Sigma \in \mathcal{S}^{+}_{p}$ and $A \in GL_{p_1,p_2}$. Then

1. $c(A\Sigma A^{\top}) = (R_2 \otimes R_1)c(\Sigma)(R_2 \otimes R_1)^{\top}$ for orthogonal $R_1 \in \mathcal{O}_{p_1}$, $R_2 \in \mathcal{O}_{p_2}$.
2. $c(A\Sigma A^{\top}) = c(\Sigma)$ if $A \in \mathcal{H}$ and $\mathcal{H}$ is a group.

Item 2 of the proposition says that if $\mathcal{H}$ is a group then c is a maximal invariant function of $\Sigma \in \mathcal{S}^{+}_{p}$ under the group action $\Sigma \mapsto H\Sigma H^{\top}$ for $H \in \mathcal{H}$, whilst the Kronecker covariance function k is an equivariant function by Proposition 2. One such group $\mathcal{H}$ is the set of Kronecker products of lower triangular matrices with positive diagonal entries, with h being the Cholesky square root. However, the results on covariance estimation in the remainder of the article are unaffected by

the choice of h as long as it is continuous, as is the case for the symmetric and Cholesky square root functions mentioned above. For notational simplicity, we use the symmetric square root function in the remaining sections of this article.

We now arrive at the main result of this section—an identifiable parametrisation of the set of covariance matrices in terms of Kronecker and core covariance matrices:

Proposition 5 The function $f : \mathcal{S}_p^+ \to \mathcal{S}_{p_1,p_2}^+ \times \mathcal{C}_{p_1,p_2}^+$ defined by $f(\Sigma) = (k(\Sigma), c(\Sigma))$ is a homeomorphism with inverse $g : \mathcal{S}_{p_1,p_2}^+ \times \mathcal{C}_{p_1,p_2}^+ \to \mathcal{S}_p^+$ given by $g(K, C) = h(K)Ch(K)^\top$.

The function g can be viewed as a parametrisation of $\mathcal{S}_p^+$ in terms of $\mathcal{S}_{p_1,p_2}^+ \times \mathcal{C}_{p_1,p_2}^+$. Conversely, the function f provides a unique representation of each $\Sigma \in \mathcal{S}_p^+$ as $\Sigma = h(K)Ch(K)^\top$ for some $K \in \mathcal{S}_{p_1,p_2}^+$ and $C \in \mathcal{C}_{p_1,p_2}^+$. We refer to this representation as KCD.

3 Core shrinkage estimation

3.1 Algebraic and geometric aspects of core shrinkage

Let $Y_1, \ldots, Y_n$ be an i.i.d. random sample from a mean-zero population of $p_1 \times p_2$ matrices with unknown covariance matrix $\Sigma \in \mathcal{S}_p^+$, where $p = p_1 \times p_2$. Letting $y_i = \mathrm{vec}(Y_i)$, we propose an estimator $\hat{\Sigma}$ of Σ obtained by shrinking the sample covariance matrix $S = \sum_{i=1}^n y_i y_i^\top / n$ towards the lower-dimensional manifold $\mathcal{S}_{p_1,p_2}^+$ of separable covariance matrices, so that our estimator $\hat{\Sigma}$ has the form

$$\hat{\Sigma} = (1 - w)S + w\hat{K}, \tag{4}$$

where $\hat{K} = k(S)$ and $w \in [0, 1]$. This estimator is equivariant with respect to transformations of the form $S \mapsto (A_2 \otimes A_1)S(A_2 \otimes A_1)^\top$ for $A_2 \otimes A_1 \in GL_{p_1,p_2}^+$, since $\hat{K} = k(S)$ and k is equivariant by Proposition 2. Note that if $w > 0$ then the estimator can be positive-definite even if n is much smaller than p, as long as n is large enough for $\hat{K}$ to be positive-definite. As mentioned in the *Introduction*, a sufficiently large n can be much smaller than p, even smaller than $p_1 \wedge p_2$.

This estimator can be motivated and understood from multiple perspectives, including an empirical Bayes perspective described in the next sub-section, and from algebraic and geometric perspectives using the KCD, as we explore here. Letting $\hat{C} = c(S)$ so that $S = \hat{K}^{1/2}\hat{C}\hat{K}^{1/2}$, we have the representation

$$\begin{aligned} \hat{\Sigma} &= (1 - w)\hat{K}^{1/2}\hat{C}\hat{K}^{1/2} + w\hat{K} \\ &= \hat{K}^{1/2}\Big[(1 - w)\hat{C} + wI_p\Big]\hat{K}^{1/2} = \hat{K}^{1/2}\hat{C}_w\hat{K}^{1/2}, \end{aligned} \tag{5}$$

where $\hat{C}_w = (1 - w)\hat{C} + wI_p$. Note that $\hat{C}_w \in \mathcal{C}_{p_1,p_2}^+$ because $\mathcal{C}_{p_1,p_2}^+$ is convex and includes I_p. The core covariance of $\hat{\Sigma}$ is $\hat{C}_w$, and the Kronecker covariance is

$$\begin{aligned} k(\hat{\Sigma}) &= \hat{K}^{1/2}k(\hat{C}_w)\hat{K}^{1/2} \\ &= \hat{K}^{1/2}I_p\hat{K}^{1/2} = \hat{K}. \end{aligned} \tag{6}$$

So, whilst $\hat{\Sigma}$ and $\hat{C}_w$ are shrunken versions of S and $\hat{C}$, respectively, $k(\hat{\Sigma})$ is equal to $k(S)$. This indicates that linear shrinkage of the sample covariance matrix S towards its Kronecker covariance $\hat{K}$ in equation (4) is equivalent to linear shrinkage of the core of S towards the identity, and so we refer to $\hat{\Sigma}$ as a core shrinkage estimator. Furthermore, the fact that $k(\hat{\Sigma}) = \hat{K}$ means that $k(\hat{\Sigma})$ is an equivariant estimator of $k(\Sigma)$, which may be a desirable property in applications where $k(\Sigma)$ is a parameter of interest.

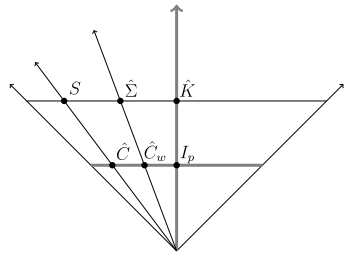

Figure 1. Core shrinkage from $S \in \mathcal{S}_p^+$ towards $\hat{K} \in \mathcal{S}_{p_1,p_2}^+$ along $\mathcal{C}_{p_1,p_2}^+$.

The geometry of core shrinkage is represented abstractly in Figure 1. The sub-manifold $\mathcal{S}_{p_1,p_2}^+$ of the cone $\mathcal{S}_p^+$ is represented by the thick grey vertical ray, and the sub-manifold $\mathcal{C}_{p_1,p_2}^+$ is represented by the thick grey horizontal line segment. For any $K \in \mathcal{S}_{p_1,p_2}^+$, the set $k^{-1}(K) = \{S : k(S) = K\}$ is known as the auxiliary space of the function k at the point K (Amari 2016, Chapter 7). The auxiliary space $k^{-1}(\hat{K})$, where $\hat{K} = k(S)$ is represented in the figure by the horizontal line that intersects $\hat{K}$, $\hat{\Sigma}$, and S. The core function c provides a bijection between each auxiliary space and the set of core covariance matrices $\mathcal{C}_{p_1,p_2}^+$. From the perspective of information geometry, the CSE shrinks S towards $\hat{K}$ along the mixture-geodesics contained in the mean parameter space of the Wishart model (Amari 1982).

3.2 Empirical Bayes estimation

Determining an appropriate amount of shrinkage to apply in equation (4) is facilitated by viewing $\hat{\Sigma}$ as an empirical Bayes estimator. Consider an inverse-Wishart prior distribution for the unknown covariance Σ,

$$\Sigma^{-1} \sim \text{Wishart}_p([(\nu - p - 1)\Sigma_2 \otimes \Sigma_1]^{-1}, \nu), \tag{7}$$

which is parametrised so that $\text{E}[\Sigma] = \Sigma_2 \otimes \Sigma_1$. The hyperparameter ν partly controls how concentrated the prior distribution of Σ is around the separable covariance matrix $\Sigma_2 \otimes \Sigma_1$. Under this prior distribution and a mean-zero normal sampling model for $Y_1, \ldots, Y_n$, the posterior distribution of Σ is

$$\Sigma^{-1} \mid S \sim \text{Wishart}_p([nS + (\nu - p - 1)\Sigma_2 \otimes \Sigma_1]^{-1}, n + \nu)$$

and the Bayes estimator under squared-error loss is the posterior mean,

$$\text{E}[\Sigma \mid S] = (1 - w)S + w\Sigma_2 \otimes \Sigma_1, \tag{8}$$

where $w = (\nu - p - 1)/(n + \nu - p - 1)$. An empirical Bayes estimator that replaces the hyperparameter $\Sigma_2 \otimes \Sigma_1$ with $\hat{K} = k(S)$ gives the estimator

$$\widehat{\text{E}[\Sigma \mid S]} = (1 - w)S + w\hat{K},$$

which is the same as in equation (4).

The amount of shrinkage w is determined by the hyperparameter ν. Our proposed empirical Bayes estimator of ν is the maximiser in ν of the marginal density $p(S \mid \nu, \Sigma_2 \otimes \Sigma_1)$ with $\hat{K}$ plugged-in for $\Sigma_2 \otimes \Sigma_1$. This density has an essentially closed-form expression due to the conjugacy of the inverse-Wishart prior distribution (7). Using standard calculations, we obtain the marginal density of S as

$$p(S \mid \nu, \Sigma_2 \otimes \Sigma_1) = r \times \frac{k(\nu)|(\nu - p - 1)\Sigma_2 \otimes \Sigma_1|^{\nu/2}}{k(\nu + n)|nS + (\nu - p - 1)\Sigma_2 \otimes \Sigma_1|^{(\nu+n)/2}},$$

where r does not depend on v and $k(v)^{-1} = 2^{vp/2}\Gamma_p(v/2)$, with Γ_p being the multivariate gamma function. Now we plug-in $\hat{K}$ for $\Sigma_2 \otimes \Sigma_1$ and utilise the fact that $S = \hat{K}^{1/2}\hat{C}\hat{K}^{1/2}$ to obtain

$$\begin{aligned} p(S \mid v, \hat{K})/b &= |\hat{K}|^{-n/2} \times \frac{k(v)}{k(v+n)}(v-p-1)^{-np/2}\left|I_p + \frac{n}{v-p-1}\hat{C}\right|^{-(v+n)/2} \\ &= \left(|\hat{K}|^{-n/2}2^{np/2}\right) \times \frac{\Gamma_p((v+n)/2)}{\Gamma_p(v/2)}(v-p-1)^{-np/2}\left|I_p + \frac{n}{v-p-1}\hat{C}\right|^{-(v+n)/2}. \end{aligned}$$

After some additional manipulation, we have that $p(S \mid v, \hat{K}) \propto_v L(v)$ where

$$L(v) = \frac{\Gamma_p((n+v)/2)}{\Gamma_p(v/2)} \times w^{vp/2}(1-w)^{np/2} \times |(1-w)\hat{C} + wI_p|^{-(v+n)/2}, \tag{9}$$

with $w = (v-p-1)/(n+v-p-1)$ as before. Computation of $L(v)$ is facilitated by noting that the determinant term can be expressed as $|(1-w)\hat{C} + wI_p| = \prod_{j=1}^{p}(w + (1-w)\hat{c}_j)$, where $\hat{c}_1, \ldots, \hat{c}_p$ are the eigenvalues of $\hat{C}$. Our proposed empirical Bayes estimator of v is the maximiser $\hat{v}$ of L, which gives $\hat{w} = (\hat{v}-p-1)/(n+\hat{v}-p-1)$ as the amount of shrinkage. The resulting empirical Bayes core shrinkage estimator is given by

$$\begin{aligned} \hat{\Sigma} &= \hat{K}^{1/2}[(1-\hat{w})\hat{C} + \hat{w}I_p]\hat{K}^{1/2} \\ &= (1-\hat{w})S + \hat{w}\hat{K}. \end{aligned} \tag{10}$$

To understand how the data influence the value of $\hat{\Sigma}$ through $\hat{w}$, write $L(v) = a(v) \times b(v)$, where $b(v) = |(1-w)\hat{C} + wI_p|^{-(v+n)/2}$ is the part of L that depends on the data, and $a(v) = L(v)/b(v)$. The function $a(v)$ is generally increasing, and so this part of L 'favours' large values of $\hat{v}$ (and $\hat{w}$). If the sample covariance S is very close to being separable, then $\hat{C}$ is very close to the identity matrix and so $b(v)$ is roughly constant in v. In this case, $a(v)$ dominates $L(v)$, resulting in a large $\hat{w}$ and strong shrinkage of S towards the sample Kronecker covariance $\hat{K}$. However, if $\hat{C}$ is far from the identity then b can be strongly decreasing in v, which results in $\hat{\Sigma}$ being close to S. In summary, the degree of shrinkage towards the space of separable covariance matrices depends on how close S is to being separable, as measured by how close $\hat{C}$ is to the identity matrix.

Finally, we note that $\hat{\Sigma}$ does not depend on the choice of separable square root function: This is because if $\hat{C}$ and $\hat{C}'$ are core matrices of S obtained from different square root functions, they still must satisfy $\hat{C}' = R\hat{C}R^\top$ for some orthogonal matrix R. This difference does not affect the empirical Bayes estimator of v, since

$$\begin{aligned} |(1-w)\hat{C}' + wI_p| &= |(1-w)R\hat{C}R^\top + wRR^\top| \\ &= |RR^\top|\,|(1-w)\hat{C} + wI_p| = |(1-w)\hat{C} + wI_p|. \end{aligned}$$

3.3 Consistency

We now provide some consistency results for the components of the KCD and the core shrinkage estimator. First, we have the very general result that a consistent estimator S of Σ can be used to obtain consistent estimators of $k(\Sigma)$ and $c(\Sigma)$, and vice versa:

Lemma 1 $k(S) \to^p k(\Sigma)$ and $c(S) \to^p c(\Sigma)$ if and only if $S \to^p \Sigma$.

This follows directly from the continuity result in Proposition 5 and the continuous mapping theorem. The lemma implies that if S is consistent and $\Sigma \in \mathcal{S}^+_{p_1,p_2}$ then the CSE $\hat{\Sigma} = (1-\hat{w})S + \hat{w}k(S)$ is consistent as well because $k(S) \to^p k(\Sigma) = \Sigma$. Conversely, if $\Sigma \notin \mathcal{S}^+_{p_1,p_2}$, then consistency

of $\hat{\Sigma}$ additionally requires that $\hat{w}$, the weight on $k(S)$, converges in probability to zero. This holds if S is consistent:

Lemma 2 If $S \to^p \Sigma$ and $\Sigma \notin \mathcal{S}^+_{p_1,p_2}$ then $\hat{w} \to^p 0$.

Lemmas 1 and 2 together give the consistency of $\hat{\Sigma}$:

Proposition 6 If $S \to^p \Sigma$ then $\hat{\Sigma} \to^p \Sigma$ for any $\Sigma \in \mathcal{S}^+_p$.

These consistency results only assume that S converges in probability to Σ as some index n, used in the definition of $\hat{w}$, goes to infinity. Even though a normal model was used to construct $\hat{w}$, the data need not be normal, and the estimator S need not be Wishart-distributed, for consistency to hold.

4 Numerical examples

4.1 Monte Carlo study

Because of its adaptive nature, we expect that the CSE $\hat{\Sigma}$ outperforms the unrestricted MLE S in general and performs nearly as well as the separable MLE $\hat{K}$ when the true covariance is exactly separable. We examine this in a finite-sample setting with a small simulation study. We considered two dimensions for the sample space, $(p_1, p_2) = (5, 7)$ and $(p_1, p_2) = (13, 17)$ which correspond to values of $p = p_1 \times p_2$ being 35 and 221, respectively. For each dimension, eight sample sizes n were considered, ranging from p_2 to $3p_1p_2/2$. Note that this includes sample sizes n that are much smaller than the dimension p. For each dimension and each sample size, population covariance matrices were generated under four scenarios, three of which were simulated from the inverse-Wishart prior distribution (7) with three values of the degrees of freedom parameter ν ranging from $p + 2$ to $3p + 1$. In the fourth scenario, which we refer to as $\nu = \infty$, Σ was set to a separable matrix (the identity matrix). To summarize, our simulation scenarios include $8 \times 4 = 32$ combinations of n and ν for each of 2 different values of (p_1, p_2).

For each of these 64 scenarios, 200 matrices Σ were simulated from equation (7), and from each a sample of n random matrices from the corresponding multivariate normal distribution were generated. From each sample, we computed four estimators: the sample covariance or MLE S, the separable MLE $\hat{K}$, the CSE $\hat{\Sigma}$, and the oracle Bayes estimator (8) which uses perfect knowledge of the hyperparameters ν and $\Sigma_2 \otimes \Sigma_1$ of the prior distribution (7). For each sample and estimator, the squared error loss in estimating Σ was computed.

Before comparing the estimators in terms of loss, we first examine the performance of the empirical Bayes estimator of $\hat{w}$ of w, which determines the amount of shrinkage towards $\hat{K}$. Results for all simulation scenarios are shown in Figure 2, where sample means of the 200 values of $\hat{w}$ are plotted as a function of the sample size. On average, $\hat{w}$ overestimates w with the bias decreasing with increasing sample size and dimension p, and also being smaller for the smaller values of w. Our intuition regarding the overestimation is that the ideal estimate of w would be obtained by evaluating how close S is to $\Sigma_2 \otimes \Sigma_1$. In contrast, $\hat{w}$ is obtained by evaluating how close S is to $\hat{K}$. Since $\hat{K}$ is the closest element of $\mathcal{S}^+_{p_1,p_2}$ to S by construction, $\hat{w}$ overestimates how close S is to $\Sigma_2 \otimes \Sigma_1$.

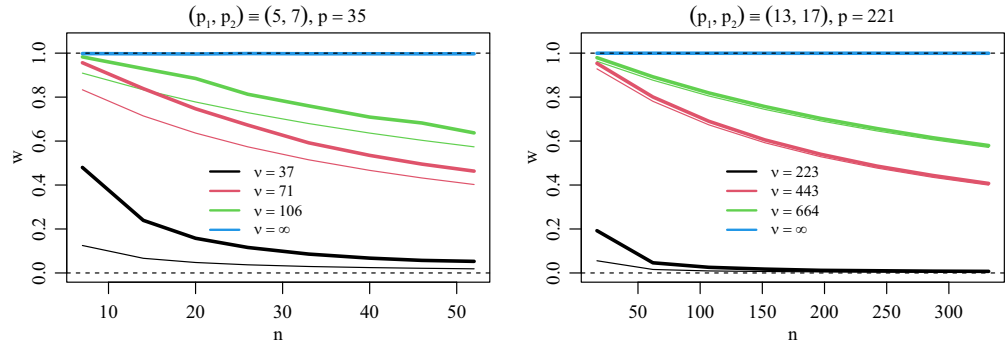

Figure 2. Average of 200 values of $\hat{w}$ as a function of ν and n in thick lines, true values of w in thin lines.

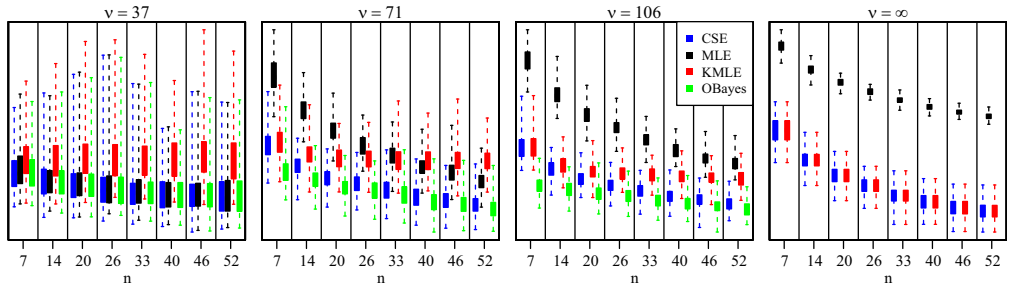

Figure 3. Loss comparisons for $(p_1, p_2) = (5, 7)$. Boxplots of the log-loss from 200 simulated data sets for each scenario. Estimators include the sample covariance matrix (maximum likelihood estimator—MLE), the separable MLE (Kronecker MLE—KMLE), the core shrinkage estimator (CSE) and the oracle Bayes estimator (OBayes).

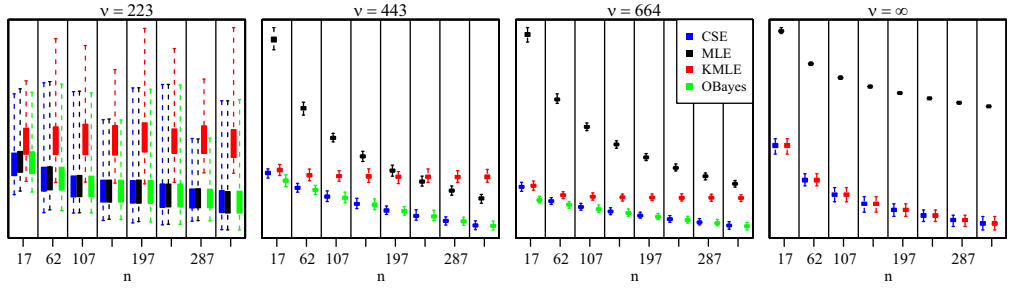

Figure 4. Loss comparisons for $(p_1, p_2) = (13, 17)$. Boxplots of the log-loss from 200 simulated data sets for each scenario. Estimators include the sample covariance matrix (MLE), the separable MLE (KMLE), the core shrinkage estimator (CSE) and the oracle Bayes estimator (OBayes).

Loss comparisons for the four estimators are displayed in Figures 3 and 4 for the $(p_1, p_2) = (5, 7)$ and $(p_1, p_2) = (13, 17)$ scenarios, respectively. The performance comparisons among the four estimators are similar in each of these two cases. The oracle Bayes estimator has the best performance for each value of ν. For the smallest values of ν, for which Σ is not close to being separable, the performance of the unrestricted MLE is nearly identical to that of the oracle Bayes estimator. This is because the value of the oracle shrinkage weight is $1/n$ and so these two estimators are nearly the same. The core shrinkage estimator (CSE) has a loss performance nearly identical to these two estimators, since for small values of ν, the estimate $\hat{\nu}$ is quite good. In contrast, the Kronecker MLE (KMLE) has worse performance on average than the other estimators, and its loss does not improve with increasing sample size. The explanation for this is that the Kronecker covariance $k(\Sigma)$ does not require a large sample size to be well estimated, and so $\hat{K}$ is close to $k(\Sigma)$ for all sample sizes, but this is far from Σ since Σ is not close to being separable.

The pattern changes somewhat for the larger values of ν. In general, the loss of the KMLE is good for small sample sizes, but does not improve much with increasing sample size since it converges to $k(\Sigma)$, which is not equal to Σ. In contrast, the unrestricted MLE is poor for small sample sizes but, since it is a consistent estimator, has a loss that steadily decreases with increasing sample size. The CSE is generally as good or better than either of these estimators across the different sample sizes: For small n it is about as good as the KMLE, and for large n, where both $k(\Sigma)$ and ν can be well estimated, it performs nearly as well as the oracle Bayes estimator.

Finally, the far-right panel of each figure gives the performance of the CSE and unrestricted and separable MLEs in the case that Σ is truly separable (the oracle Bayes estimator in this case is exactly Σ). The performance of the CSE and KMLE are nearly identical, and much better than that of the unrestricted MLE. This is not too surprising given the observation from Figure 2 that $\hat{\nu}$ tends to overestimate ν when ν is a large (finite) value. Although any finite estimate $\hat{\nu}$ of ν is in some sense too small for this case where Σ is exactly separable, $\hat{\nu}$ is generally large enough to make the shrinkage weight on w nearly equal to one, which gives an estimate that is nearly identical to the KMLE.

4.2 Speech recognition

Many data analysis tasks rely on accurate covariance estimates, including tasks that are not specifically about covariance estimation. For example, QDA is a simple and popular method of classification that relies on estimates of the population means and covariances of each potential class to which new observations are to be assigned. Specifically, the score of a new observation with feature vector $y \in \mathbb{R}^p$ with respect to category $k \in \{1, \ldots, K\}$ is

$$s_k(y) = (y - \hat{\mu}_k)^\top \hat{\Sigma}_k^{-1} (y - \hat{\mu}_k) + \ln |\hat{\Sigma}_k|,$$

where $(\hat{\mu}_k, \hat{\Sigma}_k)$ are estimates of the population mean and covariance of the feature vectors of objects in class k. If the frequencies of the different classes are equal, the classification rule is to assign the object with feature vector y to the class with the minimum score. The accuracy of such a classification procedure will depend on, among other things, the accuracy of the mean and covariance estimates for each group. In cases where the feature vector y is the vectorisation of a matrix of features, we may consider using the CSE given by equation (10) to make classifications, as an alternative to either the unstructured MLE, the separable MLE, or other types of estimators.

As a numerical illustration, we consider classification of spoken-word audio samples for 10 command words ('yes', 'no', 'up', 'down', 'left', 'right', 'on', 'off', 'stop', 'go'), using the data set provided by Warden (2017) and described in Warden (2018). The data we consider include 20,600 1-s long audio WAV files, with a per-word sample size ranging from 1,987 to 2,103 across the 10 words, representing between 989 and 1079 unique speakers for each word. We retain 100 audio samples per word for testing, and train our classifier on the remaining 19,600 audio samples. We do not make use of the fact that some speakers are represented multiple times in the data set.

A standard set of features for audio classification are mel-frequency cepstral coefficients (MFCCs), which describe an audio sample in terms of a matrix whose dimensions represent periodicities in the power spectrum of the signal across time increments. Typically, only the first 13 coefficients are retained for speech recognition tasks [Rao and Manjunath (2017, Appendix A); Sueur (2018, Chapter 12)]. For each audio sample in the data set, we computed a $p_1 \times p_2 = 99 \times 13$ matrix of the first 13 mel cepstral coefficients across 99 time bins using the function `melfcc` in the R-package `tuneR` (Ligges et al. 2018). Sample correlations for two of the words appear in Figure 5 (correlations instead of covariances are easier to visualise because of the large across-coefficient heteroscedasticity). The sample covariance matrices for these words are $p \times p = 1{,}287 \times 1{,}287$ matrices where, for example, the 99×99 block in the upper left corner is the sample covariance matrix for the first cepstral coefficient across the 99 time points.

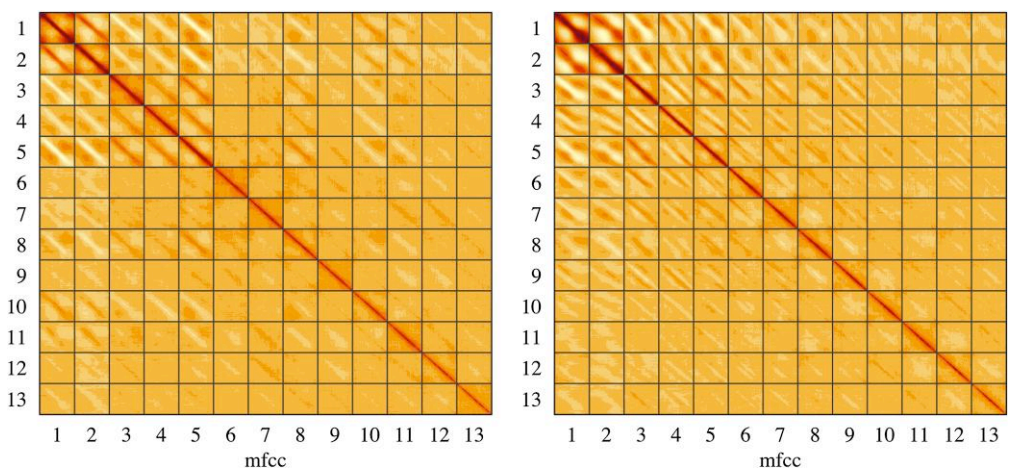

Figure 5. Correlation of mel-frequency cepstral coefficients for the word 'up' (left) and 'down' (right).

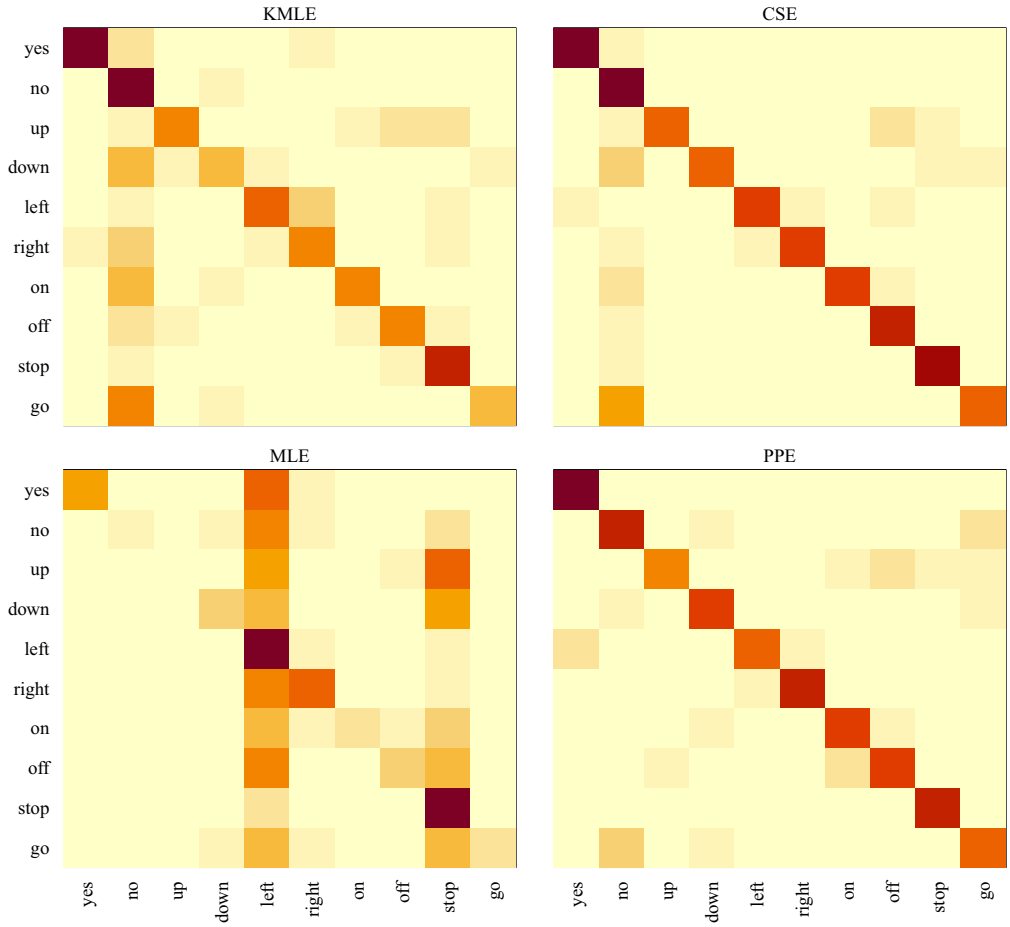

Figure 6. Confusion matrices resulting from the four covariance estimates. Rows correspond to target words and columns correspond to predictions.

From the training data, we computed sample means and several different covariance estimates for each of the 10 words. Our main interest is in comparing prediction accuracy of the core shrinkage estimator to that of the unstructured and separable MLEs, but we also compute predictions using estimates that are partially pooled across groups. Quadratic discriminant analysis with partially pooled covariance estimates often have better performance than with class-specific sample covariance matrices, particularly when the sample size n is not large compared to the dimension p. To obtain the pooling weights we use an approach outlined in Greene and Rayens (1989), which is based on an inverse-Wishart hierarchical model for $\Sigma_1, \ldots, \Sigma_{10}$. The resulting partially pooled covariance estimates (PPEs) are each roughly equal to a 32%–68% weighted average of the word-specific sample covariance and the pooled sample covariance respectively. For comparison, the estimated weights for the core shrinkage estimates (CSEs) ranged between 0.70 and 0.78 with an average of 0.75, and so these estimates were roughly 25%–75% weighted averages of the word-specific sample covariances and the word-specific Kronecker MLEs.

Classifications for the 100 training observations were made using each of the covariance estimates. Confusion matrices are displayed in Figure 6, with the true word classes along the rows, and the predicted classes along the columns. For example, the word 'go' is most frequently misclassified as 'no'. From the figure, QDA with the CSE appears to be substantially more accurate than using either the unstructured or separable MLEs, and is similar to using the partially pooled estimates. Rates of correct classification across all words for all four QDA classifiers are given in Table 1. The CSE performs better than the KMLE for all words, and better than the MLE for

Table 1. Word-specific and average rates of correct classification on the test data set for the classifiers based on the Kronecker maximum likelihood estimator (KMLE), core shrinkage estimator (CSE), unrestricted MLE (MLE), and partially pooled estimator (PPE)

	yes	no	up	down	left	right	on	off	stop	go	Average
KMLE	0.69	0.74	0.38	0.30	0.44	0.40	0.41	0.41	0.61	0.30	0.47
CSE	0.79	0.82	0.51	0.53	0.60	0.60	0.59	0.65	0.70	0.50	0.63
MLE	0.37	0.09	0.04	0.24	0.77	0.45	0.16	0.22	0.75	0.14	0.32
PPE	0.82	0.66	0.46	0.57	0.51	0.67	0.58	0.57	0.66	0.48	0.60

all words except 'left' and 'stop'. However, this apparent good performance on these two words is misleading, as it is a result of this classifier assigning most words to being either in one of these two categories, as can be seen from Figure 6. Additionally, the CSE is as good or better than the PPE for seven of the ten words. We note that the PPE is, like the CSE, a type of shrinkage estimator, although one that does not make use of the matrix structure of the data.

5 Discussion

Many classic estimators of covariance matrices are obtained by first computing the eigendecomposition of the sample covariance matrix and then regularising the resulting eigenvalues (C. Stein 1975; Takemura 1983). The CSE proposed in this article can be viewed analogously: the KCD of the sample covariance matrix is computed, and then the resulting core is regularised. However, whilst existing distributional results for the sample eigenvalues permit theoretical risk calculations for unstructured covariance estimators, we lack such detailed knowledge of the distribution of sample core matrices. Further research on the distribution of sample cores could permit theoretical comparisons of different core shrinkage estimators.

The primary computational cost of obtaining the CSE is computing the eigenvalues of the $p \times p$ sample core covariance matrix $\hat{C}$, which are needed to define the objective function (9) that determines the shrinkage weight $\hat{w}$. For very large p, the cost of this calculation may be prohibitive. For such situations, it may be possible to develop an alternative shrinkage estimator based on a partial isotropy model for the population core covariance C, that is, $C = AA^\top + I_p$ for some $A \in \mathbb{R}^{p \times r}$ with $r \ll p$. Since the space of core matrices is a convex cone that includes I_p, such a C will be a core matrix if $AA^\top$ is a core matrix. Development of such an estimator will require a characterisation of the low-rank core covariance matrices.

The results in this article extend naturally to separable covariance models for tensor-valued data, that is, data arrays having three or more index sets. For example, an empirical Bayes covariance estimator that shrinks a sample covariance matrix towards a Kronecker product of several smaller covariance matrices, one for each index set, can be derived as in Section 3.2, using the same objective function (9) to determine the amount of shrinkage. A less straightforward extension would be an estimator that adaptively shrinks towards an appropriate separable submodel, that is, a submodel that is separable after some groups of indices of the data array have been collapsed.

Data availability

Replication code for the numerical results in this article are available from the R-package `covKCD` and the corresponding repository github.com/pdhoff/covKCD.

Conflict of interest: None declared.

Appendix A. Proofs

Proof of Proposition 1. We first obtain an identity that relates the expectations in equation (2) to the trace term in the divergence function $d(K:\Sigma)$. Letting y be the vectorisation of Y, for $(K_1, K_2) \in \mathcal{S}^+_{p_1} \times \mathcal{S}^+_{p_2}$ we have

$$\begin{aligned}\operatorname{trace}((K_2^{-1} \otimes K_1^{-1})\Sigma) &= \mathrm{E}[\operatorname{trace}((K_2^{-1} \otimes K_1^{-1})yy^\top)] \\ &= \mathrm{E}[y^\top(K_2^{-1} \otimes K_1^{-1})y] \\ &= \mathrm{E}[\operatorname{trace}(Y^\top K_1^{-1} Y K_2^{-1})] \\ &= \operatorname{trace}(K_1^{-1}\mathrm{E}[YK_2^{-1}Y^\top]) = \operatorname{trace}(K_2^{-1}\mathrm{E}[Y^\top K_1^{-1}Y]).\end{aligned}$$

Therefore, for $K = K_2 \otimes K_1$ the divergence function may be written

$$\begin{aligned}d(K:\Sigma) &= p_2 \ln|K_1| + p_1 \ln|K_2| + \operatorname{trace}(K_1^{-1}\mathrm{E}[YK_2^{-1}Y^\top]) \\ &= p_2 \ln|K_1| + p_1 \ln|K_2| + \operatorname{trace}(K_2^{-1}\mathrm{E}[Y^\top K_1^{-1}Y]).\end{aligned}$$

Now suppose that $\Sigma_2 \otimes \Sigma_1 \in \mathcal{S}^+_{p_1,p_2}$ minimises the divergence. Then Σ_1 must also be the minimiser of the divergence in K_1 when K_2 is fixed at Σ_2, that is, Σ_1 minimises $p_2 \ln|K_1| + \operatorname{trace}(K_1^{-1}\mathrm{E}[Y\Sigma_2^{-1}Y^\top])$ over $K_1 \in \mathcal{S}^+_{p_1}$. It is well known (Anderson 2003, Section 4.1) that this function of K_1 is uniquely minimised by $\mathrm{E}[Y\Sigma_2^{-1}Y^\top]/p_2$, and so $\Sigma_1 = \mathrm{E}[Y\Sigma_2^{-1}Y^\top]/p_2$. Similarly, Σ_2 must equal $\mathrm{E}[Y^\top\Sigma_1^{-1}Y]/p_1$, and so (Σ_1, Σ_2) is a solution to equation (2).

Conversely, let $f(K_1, K_2:\Sigma) = \ln|K_2 \otimes K_1| + \operatorname{trace}(K_2 \otimes K_1)^{-1}\Sigma))$ be the divergence written as a real-valued function on $\mathcal{S}_{p_1} \times \mathcal{S}_{p_2}$. Differentiating f with respect to (K_1, K_2) shows that the stationary points of f are the solutions to (2). Although f is not convex, it is geodesically convex (Wiesel 2012), and so by Corollary 3.1 of Rapcsák (1991), every stationary point of f is a global minimiser of f. Thus, if (K_1, K_2) is a solution to (2) then $K_2 \otimes K_1$ is a minimiser of d. □

Proof of Proposition 2. Let $A = A_2 \otimes A_1$. For each K, we have

$$\begin{aligned}d(K:A\Sigma A^\top) &= \ln|K| + \operatorname{trace}(K^{-1}A\Sigma A^\top) \\ &= \ln|A^{-1}KA^{-\top}| + \operatorname{trace}((A^{-1}KA^{-\top})^{-1}\Sigma) + \ln|AA^\top| \\ &\equiv d(\tilde{K}:\Sigma) + \ln|AA^\top|,\end{aligned}$$

where $\tilde{K} = A^{-1}KA^{-\top}$. Note that for $A \in GL_{p_1,p_2}$, $\{A^{-1}KA^{-\top} : K \in \mathcal{S}^+_{p_1,p_2}\} = \mathcal{S}^+_{p_1,p_2}$. By Proposition 1, $d(\tilde{K}:\Sigma)$ is minimised by $\tilde{K} = \Sigma_2 \otimes \Sigma_1$, and so $d(K:A\Sigma A^\top)$ is minimised by $K = A\tilde{K}A^\top = (A_2\Sigma_2A_2^\top) \otimes (A_1\Sigma_1A_1^\top)$. □

Proof of Corollary 1. Items 1 and 2 can be shown by noting that the unconstrained minimiser of $\ln|K| + \operatorname{trace}(K^{-1}\Sigma)$ over $K \in \mathcal{S}^+_p$ is Σ, and so if $\Sigma \in \mathcal{S}^+_{p_1,p_2}$ then the minimiser over $K \in \mathcal{S}^+_{p_1,p_2}$ is Σ as well. Alternatively, Item 1 can be shown by noting that $I_p = I_{p_2} \otimes I_{p_1}$, and confirming that

(I_{p_1}, I_{p_2}) provide a solution to equation (2) when $\mathrm{Var}[Y] = I_p$. Item 2 can also be shown this way, or with Proposition 2: If $\Sigma = \Sigma_2 \otimes \Sigma_1$ then

$$\begin{aligned} k(\Sigma) &= k((\Sigma_2^{1/2} \otimes \Sigma_1^{1/2}) I_p (\Sigma_2^{1/2} \otimes \Sigma_1^{1/2})) \\ &= (\Sigma_2^{1/2} \otimes \Sigma_1^{1/2}) k(I_p) (\Sigma_2^{1/2} \otimes \Sigma_1^{1/2}) \\ &= (\Sigma_2^{1/2} \otimes \Sigma_1^{1/2})(\Sigma_2^{1/2} \otimes \Sigma_1^{1/2}) = \Sigma. \end{aligned}$$

Item 3 can also be obtained from Proposition 2 by choosing (for example) $A_1 = aI_{p_1}$ and $A_2 = I_{p_2}$. Finally, if $\mathrm{Var}[Y] = \Sigma$ is diagonal then $\mathrm{E}[y_i^\top A_1 y_{i'}] = 0$ for rows y_i and $y_{i'}$ of Y for any matrix $A_1 \in \mathbb{R}^{p_2 \times p_2}$ unless $i = i'$. As a result, $\mathrm{E}[Y A_1 Y^\top]$ is diagonal, as is $\mathrm{E}[Y^\top A_2 Y^\top]$ for the same reason. This implies that if (Σ_1, Σ_2) is a solution to equation (2) then both matrices are diagonal, as is their Kronecker product. □

Proof of Proposition 3. If $\mathrm{E}[YY^\top]/p_2 = I_{p_1}$ and $\mathrm{E}[Y^\top Y]/p_1 = I_{p_2}$ then (I_{p_1}, I_{p_2}) is a solution to equation (2) and so $k(C) = I_{p_2} \otimes I_{p_1} = I_p$. Conversely, if $k(C) = I_p$ then any solution to equation (2) is of the form $(cI_{p_1}, c^{-1} I_{p_2})$ for some $c > 0$, which implies $\mathrm{E}[YY^\top]/p_2 = I_{p_1}$ and $\mathrm{E}[Y^\top Y]/p_1 = I_{p_2}$. Finally, let y_j be the jth column vector of Y. Then

$$\mathrm{E}[YY^\top] = \sum_{j=1}^{p_2} \mathrm{E}[y_j y_j^\top] = \sum_{j=1}^{p_2} \tilde{C}_{\cdot,j,\cdot,j}.$$

□

Proof of Proposition 4. Let $c(\Sigma) = C$, $k(\Sigma) = K$ and $h(K) = H$, so $\Sigma = HCH^\top$. By Proposition 2, $k(A\Sigma A^\top) = AKA^\top = AHH^\top A^\top$. Let $\tilde{K} = AKA^\top$ and $\tilde{H} = h(\tilde{K})$. Then $c(A\Sigma A^\top) = \tilde{H}^{-1}(AH)C(AH)^\top \tilde{H}^{-\top}$. But by the definition of the square root function, we must have $\tilde{H}\tilde{H}^\top = \tilde{K} = AHH^\top A^\top$, and so $\tilde{H} = AHR^\top$ for some $R \in \mathcal{O}_p$. Furthermore this R must be separable because both $\tilde{H}$ and AH are separable. Thus $\tilde{H}^{-1} = RH^{-1}A^{-1}$ and Item 1 of the result follows. If $A \in \mathcal{H}$ and $\mathcal{H}$ is a group, then $AH \in \mathcal{H}$, and so $\tilde{H} \equiv h(AHH^\top A^\top) = AH$, giving Item 2. □

Proof of Proposition 5. First, we show that f is a bijection. For any $\Sigma \in \mathcal{S}_p^+$, let $H = h(k(\Sigma))$ and $C = c(\Sigma)$. Then

$$\begin{aligned} g(f(\Sigma)) &= HCH^\top \\ &= H(H^{-1}\Sigma H^{-\top})H^\top = \Sigma. \end{aligned}$$

Conversely, let $(C, K) \in \mathcal{C}^+_{p_1,p_2} \times \mathcal{S}^+_{p_1,p_2}$. Then with $H = h(K)$, we have

$$\begin{aligned} f(g(C, K)) &= f(HCH^\top) \\ &= (k(HCH^\top), c(HCH^\top)). \end{aligned}$$

Since $H \in \mathcal{S}^+_{p_1,p_2}$, by Proposition 2 we have

$$\begin{aligned} k(HCH^\top) &= Hk(C)H^\top \\ &= HIH^\top \\ &= HH^\top = K. \end{aligned}$$

Finally,

$$\begin{aligned} c(HCH^\top) &= h(K)^{-1}(HCH^\top)h(K)^{-\top} \\ &= H^{-1}(HCH^\top)H^{-\top} = C, \end{aligned}$$

and so $f(g(C, K)) = (C, K)$.

We now show that the Kronecker covariance function k is continuous, from which the continuity results for f and g follow. The space $\mathcal{S}^+_p$ is a complete Riemannian manifold with respect to the affine invariant metric $d_A : \mathcal{S}^+_p \times \mathcal{S}^+_p \to \mathbb{R}^+$ given by

$$d_A(\Sigma, \tilde{\Sigma}) = \| \log(\Sigma^{-1/2}\tilde{\Sigma}\Sigma^{-1/2}) \|,$$

where 'log' is the matrix logarithm (Bhatia 2007; Higham 2008). Note that by the form of d_A and the fact that $d_A(\Sigma, \tilde{\Sigma}) \leq d_A(\Sigma, I_p) + d_A(I_p, \tilde{\Sigma})$, a subset of $\mathcal{S}^+_p$ is bounded under this metric if and only if the eigenvalues of its elements are bounded away from zero and infinity.

Let $\{S_n\}$ be a sequence in $\mathcal{S}^+_p$ that converges to $\Sigma \in \mathcal{S}_{p^+}$ in this metric. Convergence of the sequence implies it is bounded, and so there exists an interval $[a, b] \subset (0, \infty)$ that contains the eigenvalues of S_n for all n. We now show that boundedness of $\{S_n\}$ implies that the sequence $\{K_n\} \equiv \{k(S_n)\}$ is bounded. Recall that K_n is the minimiser of the divergence d over $\mathcal{S}^+_{p_1,p_2}$, and so $d(K_n : S_n) \leq d(I_p : S_n)$. Using this fact and the bounds on the eigenvalues of $\{S_n\}$, we have

$$\sum_{j=1}^{p} (\log l_{n,j} + a/l_{n,j}) = d(K_n : aI_p) \leq d(K_n : S_n) \leq d(I_p : S_n) \leq pb,$$

where $l_{n,j}$ is the jth largest eigenvalue of S_n. Noting that $\log x + a/x$ is a convex function with a minimum at $x = a$, we have for each $k \in \{1, \ldots, p\}$

$$\begin{aligned} \log l_{n,k} + a/l_{n,k} &\leq pb - \sum_{j \neq k} (\log l_{n,j} + a/l_{n,j}) \\ &\leq pb - (p-1) \times (\log a + 1). \end{aligned}$$

Since $\log x + a/x$ diverges as x goes to zero or infinity, the above bound implies that there exists $[c, d] \subset (0, \infty)$ that contains $l_{n,k}$ for all n and k, that is, $\{K_n\}$ is bounded.

Now let $\{K_{n_s}\}$ be any convergent subsequence of $\{K_n\}$ and let $K = k(\Sigma)$. Let $K_{n_s} \to K^*$, and so $d(K_{n_s} : S_{n_s}) \leq d(K : S_{n_s})$. Since d is jointly continuous in both of its arguments, taking the limit of the previous inequality gives $d(K^* : \Sigma) \leq d(K : \Sigma)$, which implies that $K^* = K$. This implies that $K_n \to K$ because the closure of

the bounded set $\{K_n\}$ is itself bounded and therefore sequentially compact by the completeness of $\mathcal{S}_p^+$. Thus k is continuous. Furthermore, since the topology of $\mathcal{S}_p^+$ under the affine invariant metric is the same as that under the Euclidean metric (Lee 2018, Theorem 2.55), k is continuous for this metric space as well. Finally, the functions f and g are continuous because they are both compositions of the continuous function k with other continuous functions. □

Proof of Lemma 2. We first find a limiting form for an objective function from which $\hat{v}$ is obtained. To facilitate our analysis, we use the objective function $l_n(r, \hat{C}) = -2 \log L(nr)/n + p(\log \frac{n}{2} - 1)$ with L defined in equation (9), so that the estimated value of v is $\hat{v} = n \times \arg\min_{r \geq (p+1)/n} l_n(r, \hat{C}) = \arg\max_{v \geq p+1} L(v)$, where now we make explicit the dependence of the objective function on the sample core matrix $\hat{C}$. As a function of $(r, C) \in \mathbb{R}^+ \times \mathcal{C}^+_{p_1,p_2}$, the objective function is then $l_n(r, C) = a_n + b_n(r) + c_n(r, C)$ where $a_n = p(\log \frac{n}{2} - 1)$ and

$$b_n = -\frac{2}{n} \log\left(\frac{\Gamma_p(n(1+r)/2)}{\Gamma_p(nr/2)}\right) + p(1+r)\log(1+r+\delta) - pr\log(r+\delta)$$

$$c_n = (1+r)\log|(1-w)C + wI_p|,$$

where $w = (r+\delta)/(1+r+\delta)$ with $\delta = -(p+1)/n$. We will show that as $n \to \infty$, $a_n + b_n(r)$ converges uniformly to zero for $r \in [\epsilon, \infty)$ and $c_n(r, C)$ converges uniformly to $l(r, C)$, where

$$l(r, C) = (1+r)\log|C/(1+r) + rI_p/(1+r)|.$$

We start by showing convergence of $c_n(r, C)$ to $l(r, C)$, i.e. that the difference between w and $r/(1+r)$ is asymptotically negligible. To see this, recall that the log determinant of a matrix is a continuous function, and so is uniformly continuous on the compact set of convex combinations of core matrices and the identity. Next, we have that $(1-w)C + wI_p$ converges uniformly to $C/(1+r) + rI_p/(1+r)$, because the norm of their difference is $\|(w - \frac{r}{1+r})(I_p - C)\| < \sqrt{p(p-1)}|w - \frac{r}{r+1}|$, and $|w - \frac{r}{1+r}| = \frac{\delta}{(1+r+\delta)(1+r)}$ converges to zero uniformly in r for $r > 0$.

Next we use Stirling's approximation $\log(\Gamma(z)) = z\log(z) - z + \frac{1}{2}\log(2\pi/z) + O(z^{-1})$ on the multivariate gamma terms of $b_n(r)$. Letting $\delta_j = (1-j)/n$, we have

$$-\frac{2}{n}\left(\log\left(\Gamma_p(n(1+r)/2)\right) - \log\left(\Gamma_p(nr/2)\right)\right)$$

$$= -\frac{2}{n}\sum_{j=1}^{p} \log\left(\Gamma\left(\frac{n(1+r)+1-j}{2}\right)\right) + \log\left(\Gamma\left(\frac{nr+1-j}{2}\right)\right)$$

$$= \sum_{j=1}^{p}\left(-(1+r+\delta_j)\log\left(n(1+r+\delta_j)/2\right) + (r+\delta_j)\log\left(n(r+\delta_j)/2\right) + 1\right)$$

$$-\frac{1}{n}\sum_{j=1}^{p}\log\left(\frac{r+\delta_j}{1+r+\delta_j}\right) + O\left(n^{-1}(1+r)^{-1}\right) + O\left((nr)^{-1}\right).$$

The last three terms in the above expression converge uniformly to 0 over $r \in [\epsilon, \infty)$ for any $\epsilon > 0$. Adding a_n and the remaining terms of $b_n(r)$ gives $a_n + b_n(r)$ being approximately equal to

$$-r\sum_{j=1}^{p}\log\left(\frac{(1+r+\delta_j)(r+\delta)}{(1+r+\delta)(r+\delta_j)}\right) - \sum_{j=1}^{p}\log\left(\frac{1+r+\delta_j}{1+r+\delta}\right) - \sum_{j=1}^{p}\delta_j\log\left(\frac{1+r+\delta_j}{r+\delta_j}\right).$$

The second and third sums above converge uniformly to zero over $r \in [\epsilon, \infty)$. Regarding the first sum, consider the ratio

$$\left(\frac{1+r+\delta_j}{1+r+\delta}\right)^r = \left(1+\frac{\delta_j-\delta}{1+r+\delta}\right)^{r+1+\delta}\left(1+\frac{\delta_j-\delta}{1+r+\delta}\right)^{-(1+\delta)}.$$

The log of the second factor on the right converges to zero uniformly in r. For the first factor we have

$$1 \le \left(1+\frac{\delta_j-\delta}{1+r+\delta}\right)^{1+r+\delta} \le e^{|\delta_j|+|\delta|} \to 1$$

as $n \to \infty$, where the first inequality follows from $\delta_j - \delta \ge 0$. Similarly,

$$1 \ge \left(\frac{r+\delta}{r+\delta_j}\right)^r = \left(1+\frac{\delta-\delta_j}{r+\delta_j}\right)^r \ge \left(1+\frac{\delta-\delta_j}{\epsilon+\delta_j}\right)^{\epsilon+\delta_j}\left(1+\frac{\delta-\delta_j}{r+\delta_j}\right)^{-\delta_j} \to 1$$

as $n \to \infty$. Thus, $a_n + b_n(r)$ converges uniformly to zero on $r \in [\epsilon, \infty)$ for any $\epsilon > 0$.

The above calculation shows that our objective function $l_n(r, C)$ converges uniformly to $l(r, C)$ for $(r, C) \in [\epsilon, \infty) \times \mathcal{C}^+_{p_1,p_2}$. We want to show that this limiting objective function is strictly increasing in r if $C \neq I_p$, so in this scenario where Σ is not separable the estimated weight on the sample Kronecker covariance converges to zero. To see that this is the case, let $c_1, \ldots, c_p$ be the eigenvalues of C, so that

$$l(r, C) = \sum_{j=1}^{p}(1+r)\log\left(\frac{r+c_j}{1+r}\right) = \sum_{j=1}^{p}(1+r)\log\left(1+\frac{c_j-1}{1+r}\right).$$

The derivative of the jth term of the sum with respect to r is $\log(1+\frac{c_j-1}{1+r}) - \frac{c_j-1}{r+c_j}$. Since $\log(1+x) \ge x/(1+x)$ for $x > -1$ (with strict inequality for $x \neq 0$) this derivative is positive for $(c_j-1)/(r+1) > -1$, or equivalently, for $c_j > -r$, which holds for each $j = 1, \ldots, p$ because C is positive-definite. Additionally, because $C \neq I_p$ there is at least one j for which $c_j \neq 1$, so at least one term in the sum has a strictly positive derivative, making our objective function a strictly increasing function of r.

Finally, let $\hat{C} = c(S)$ and $C = c(\Sigma)$. We want to show that $\hat{r}$, the minimiser of $l_n(r, \hat{C})$ over $r \ge (p+1)/n$, converges in probability to zero if $C \neq I_p$, or equivalently $\Pr(\hat{r} > \epsilon) \to 0$ for any $\epsilon > 0$. By the result in the previous paragraph, $l(\epsilon/2, C) < l(\epsilon, C)$ and by the continuity of l there

is a ball B around C that does not contain I_p such that

$$\inf_{\tilde{C}\in B} l(\epsilon, \tilde{C}) - \sup_{\tilde{C}\in B} l(\epsilon/2, \tilde{C}) = \delta > 0.$$

By the uniform convergence of l_n to l, there is an N such that $|l_n(r, \tilde{C}) - l(r, \tilde{C})| < \delta/2$ for $n > N$ and all $\tilde{C} \in B$ and $r \geq \epsilon/2$. If $\hat{C} \in B$ then for any $r \geq \epsilon$

$$\begin{aligned} l_n(\epsilon/2, \hat{C}) < l(\epsilon/2, \hat{C}) + \delta/2 &\leq \sup_{\tilde{C}\in B} l(\epsilon/2, \tilde{C}) + \delta/2 \\ &= \inf_{\tilde{C}\in B} l(\epsilon, \tilde{C}) - \delta/2 \\ &\leq l(\epsilon, \hat{C}) - \delta/2 \\ &\leq l(r, \hat{C}) - \delta/2 < l_n(r, \hat{C}), \end{aligned}$$

and so $\hat{r} < \epsilon$ for $n > N$ and $\hat{C} \in B$. Thus, $\Pr(\hat{r} > \epsilon) \leq \Pr(\hat{C} \notin B) \to 0$ as $n \to \infty$, because B is a neighbourhood of C and $\hat{C}$ is consistent for C by Lemma 1. Thus, $\hat{r}$ and $\hat{w}$ converge in probability to zero as $n \to \infty$ if $C \neq I$. □

References

Amari S.-I. (1982). Differential geometry of curved exponential families—Curvatures and information loss. *The Annals of Statistics*, *10*(2), 357–385. https://doi.org/10.1214/aos/1176345779

Amari S.-I. (2016). *Information geometry and its applications*. Applied Mathematical Sciences. (Vol. 194). Springer.

Anderson T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics (3rd ed.). Wiley-Interscience (John Wiley & Sons).

Bhatia R. (2007). *Positive definite matrices*. Princeton Series in Applied Mathematics. Princeton University Press.

Dawid A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, *68*(1), 265–274. https://doi.org/10.1093/biomet/68.1.265

Derksen H., & Makam V. (2021). Maximum likelihood estimation for matrix normal models via quiver representations. *SIAM Journal on Applied Algebra and Geometry*, *5*(2), 338–365. https://doi.org/10.1137/20M1369348

Drton M., Kuriki S., & Hoff P. (2021). Existence and uniqueness of the Kronecker covariance MLE. *The Annals of Statistics*, *49*(5), 2721–2754. https://doi.org/10.1214/21-AOS2052

Dutilleul P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, *64*, 2105–123. https://doi.org/10.1080/00949659908811970

Gerard D., & Hoff P. (2015). Equivariant minimax dominators of the MLE in the array normal model. *Journal of Multivariate Analysis*, *137*, 32–49. https://doi.org/10.1016/j.jmva.2015.01.020

Gerard D., & Hoff P. (2016). A higher-order LQ decomposition for separable covariance models. *Linear Algebra and its Applications*, *505*, 57–84. https://doi.org/10.1016/j.laa.2016.04.033

Greene T., & Rayens W. S. (1989). Partially pooled covariance matrix estimation in discriminant analysis. *Communications in Statistics—Theory and Methods*, *18*(10), 3679–3702. https://doi.org/10.1080/03610928908830117

Greenewald K., Zelnio E., & Hero A. H. (2016). Robust SAR STAP via Kronecker decomposition. *IEEE Transactions on Aerospace and Electronic Systems*, *52*(6), 2612–2625. https://doi.org/10.1109/TAES.2016.150712

Higham N. J. (2008). *Functions of matrices*. Society for Industrial and Applied Mathematics (SIAM).

Hoff P. D. (2016). Limitations on detecting row covariance in the presence of column covariance. *Journal of Multivariate Analysis*, *152*, 249–258. https://doi.org/10.1016/j.jmva.2016.09.003

Huber P. (1967). The behavior of maximum likelihood estimators under non-standard conditions. In L. LeCam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 221–233). University of California Press.

Lee J. M. (2018). *Introduction to Riemannian manifolds*. Graduate Texts in Mathematics (Vol. 176). Springer.

Ligges U., Krey S., Mersmann O., & Schnackenberg S. (2018). tuneR: Analysis of music and speech.

Mardia K. V., & Goodall C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate environmental statistics*. North-Holland Series in Statistics and Probability (Vol. 6, pp. 347–386). North-Holland.

Masak T., & Panaretos V. M. (2022). Random surface covariance estimation by shifted partial tracing. *Journal of the American Statistical Association*, 1–13. https://doi.org/10.1080/01621459.2022.2061982

Masak T., Sarkar S., & Panaretos V. M. (2023). Separable expansions for covariance estimation via the partial inner product. *Biometrika*, *110*(1), 225–247. https://doi.org/10.1093/biomet/asac035

Rao K. S., & Manjunath K. E. (2017). *Speech recognition using articulatory and excitation source features*. Springer.

Rapcsák T. (1991). Geodesic convexity in nonlinear optimization. *Journal of Optimization Theory and Applications*, *69*(1), 169–183.

Roś B., Bijma F., de Munck J. C., & de Gunst M. C. M. (2016). Existence and uniqueness of the maximum likelihood estimator for models with a Kronecker product covariance structure. *Journal of Multivariate Analysis*, *143*, 345–361. https://doi.org/10.1016/j.jmva.2015.05.019

Rougier J. (2017). A representation theorem for stochastic processes with separable covariance functions, and its implications for emulation. https://arxiv.org/abs/1702.05599

Soloveychik I., & Trushin D. (2016). Gaussian and robust Kronecker product covariance estimation: Existence and uniqueness. *Journal of Multivariate Analysis*, *149*, 92–113. https://doi.org/10.1016/j.jmva.2016.04.001

Srivastava M. S., von Rosen T., & von Rosen D. (2008). Models with a Kronecker product covariance structure: Estimation and testing. *Mathematical Methods of Statistics*, *17*(4), 357–370. https://doi.org/10.3103/S1066530708040066

Stein C. (1975). Estimation of a covariance matrix. In Rietz Lecture, 39th Annual Meeting of the IMS, Atlanta, GA.

Stein M. L. (2005). Space-time covariance functions. *Journal of the American Statistical Association*, *100*(469), 310–321. https://doi.org/10.1198/016214504000000854

Sueur J. (2018). *Sound analysis and synthesis with R*. Springer.

Takemura A. (1983). *An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population* (Technical Report). DTIC Document.

Warden P. (2017). *Speech commands dataset streaming test version 2*. http://download.tensorflow.org/data/speech_commands_streaming_test_v0.02.tar.g.

Warden P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition.

Werner K., Jansson M., & Stoica P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, *56*(2), 478–491. https://doi.org/10.1109/TSP.2007.907834

Wiesel A. (2012). Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, *60*(12), 6182–6189. https://doi.org/10.1109/TSP.2012.2218241

Yin J., & Li H. (2012). Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis*, *107*, 119–140. https://doi.org/10.1016/j.jmva.2012.01.005

Zhang Y., & Schneider J. (2010). Learning multiple tasks with a sparse matrix-normal penalty. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. *23*). Curran Associates, Inc.